

# Lexical Semantic Change: Models, Data and Evaluation



LREC 2022 - Tutorial - 20 June 2022

Pierpaolo Basile<sup>1</sup>, Annalina Caputo<sup>2</sup>, Pierluigi Cassotti<sup>1</sup> and Rossella Varvara<sup>3</sup>  
University of Bari<sup>1</sup>, Dublin City University<sup>2</sup>, Université de Fribourg<sup>3</sup>

**Data**

**Diachronic resources**



# Diachronic corpora

# The New York Times Annotated Corpus

- Contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007
  - with article metadata
- The corpus provides:
  - over 1.8 million articles
  - over 650,000 article summaries
  - over 1,500,000 articles manually tagged with tags drawn from a normalized indexing vocabulary of people, organizations, locations and topic descriptors
  - over 275,000 algorithmically-tagged articles that have been hand verified
  - Java tools for parsing corpus XML documents
- Text is formatted according to the News Industry Text Format (NITF)

# Corpus of Contemporary American English

- Monitor corpus for American English
- Time period: 1990-2009
- 400 million words corpus is evenly divided between spoken, fiction, popular magazines, newspapers, and academic journals.
- The genre balance stays almost exactly the same from year to year

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4), 447-464.

# Corpus of Historical American English (COHA)

- Time period: 1810 - 2009
- Genres: newspapers, popular magazines, fiction and nonfiction books
- Size: 406 million words and around 107,000 texts.
- Balanced by genre, sub-genre and domain across decades.

# Corpus of Historical American English (COHA)

Decade	Fiction	Magazines	Newspaper	NF Books	Total	Percent fiction
1810s	641,164	88,316	0	451,542	1,181,022	0.54
1820s	3,751,204	1,714,789	0	1,461,012	6,927,005	0.54
1830s	7,590,350	3,145,575	0	3,038,062	13,773,987	0.55
1840s	8,850,886	3,554,534	0	3,641,434	16,046,854	0.55
1850s	9,094,346	4,220,558	0	3,178,922	16,493,826	0.55
1860s	9,450,562	4,437,941	262,198	2,974,401	17,125,102	0.55
1870s	10,291,968	4,452,192	1,030,560	2,835,440	18,610,160	0.55
1880s	11,215,065	4,481,568	1,355,456	3,820,766	20,872,855	0.54
1890s	11,212,219	4,679,486	1,383,948	3,907,730	21,183,383	0.53
1900s	12,029,439	5,062,650	1,433,576	4,015,567	22,541,232	0.53
1910s	11,935,701	5,694,710	1,489,942	3,534,899	22,655,252	0.53
1920s	12,539,681	5,841,678	3,552,699	3,698,353	25,632,411	0.49
1930s	11,876,996	5,910,095	3,545,527	3,080,629	24,413,247	0.49
1940s	11,946,743	5,644,216	3,497,509	3,056,010	24,144,478	0.49
1950s	11,986,437	5,796,823	3,522,545	3,092,375	24,398,180	0.49
1960s	11,578,880	5,803,276	3,404,244	3,141,582	23,927,982	0.48
1970s	11,626,911	5,755,537	3,383,924	3,002,933	23,769,305	0.49
1980s	12,152,603	5,804,320	4,113,254	3,108,775	25,178,952	0.48
1990s	13,272,162	7,440,305	4,060,570	3,104,303	27,877,340	0.48
2000s	14,590,078	7,678,830	4,088,704	3,121,839	29,479,451	0.49
Total	207,633,395	97,207,399	40,124,656	61,266,574	<b>406,232,024</b>	0.51

# Corpus of Historical American English (COHA)

	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
<i>to_v</i>	86	129	156	178	281	383	437	419	346	372	323	338	288	300	400
<i>V-ing</i>	1	8	13	12	22	30	33	49	49	54	60	77	109	138	245
<i>% V-ing</i>	0.01	0.06	0.08	0.06	0.07	0.07	0.07	0.10	0.12	<b>0.13</b>	<b>0.16</b>	<b>0.19</b>	<b>0.27</b>	<b>0.32</b>	<b>0.38</b>



# Bank of English (Cobuild Corpus)

Bank of English (BoE), also known as the Cobuild Corpus, and (in its most recent incarnation online) as Word Banks Online (<http://wordbanks.harpercollins.co.uk>).

- Based on the Collins Cobuild dictionaries
- Time period: 1980
- 455 million words by 2005

Time period	Fiction	Total	% fiction
1960–79	1,030,000	1,414,000	72.8%
1980–89	3,087,000	8,792,000	35.1%
<b>1990–94</b>	6,049,000	20,833,000	<b>29.0%</b>
<b>1995–99</b>	3,100,000	19,187,000	<b>16.2%</b>
2000–4	18,800,000	123,055,000	15.3%

# Other English corpora

- Brown family (Brown, LOB, FROWN, FLOB, Hundt and Leech, 2012):  
1960s-1990s;  
four million words.
- ARCHER Corpus (Biber *et al.*, 1994; Yáñez-Bouza 2011):  
1700s - 1900s;  
two million words
- Diachronic Corpus of Present-day Spoken English or DCPSE (Davies, 2009b):  
1950s - 1990s  
british english  
less than one million words

# Google N-gram Viewer

- Search and visualize **n-gram statistics** from Google Books
- **N-gram**: sequence of n words
- Google Books digitalizes **millions of books**

# Google N-gram

*“Google Books digitalizes millions of books”*

## 1-gram

Google, Books, digitalizes, millions, of, books

## 2-gram

Google Books, Books digitalizes, digitalizes millions, millions of, of books

## 3-gram

Google Books digitalizes, Books digitalizes millions, digitalizes millions of, millions of books

# Google N-gram: Data Format

ngram TAB year TAB match\_count TAB volume\_count NEWLINE

Example:

circumvallate	1978	335	91
circumvallate	1979	261	91

Each language for each n-gram has several .gz archives with one n-gram for each line.

Download: <https://storage.googleapis.com/books/ngrams/books/datasetsv3.html>

# Google N-gram Viewer



<https://books.google.com/ngrams>

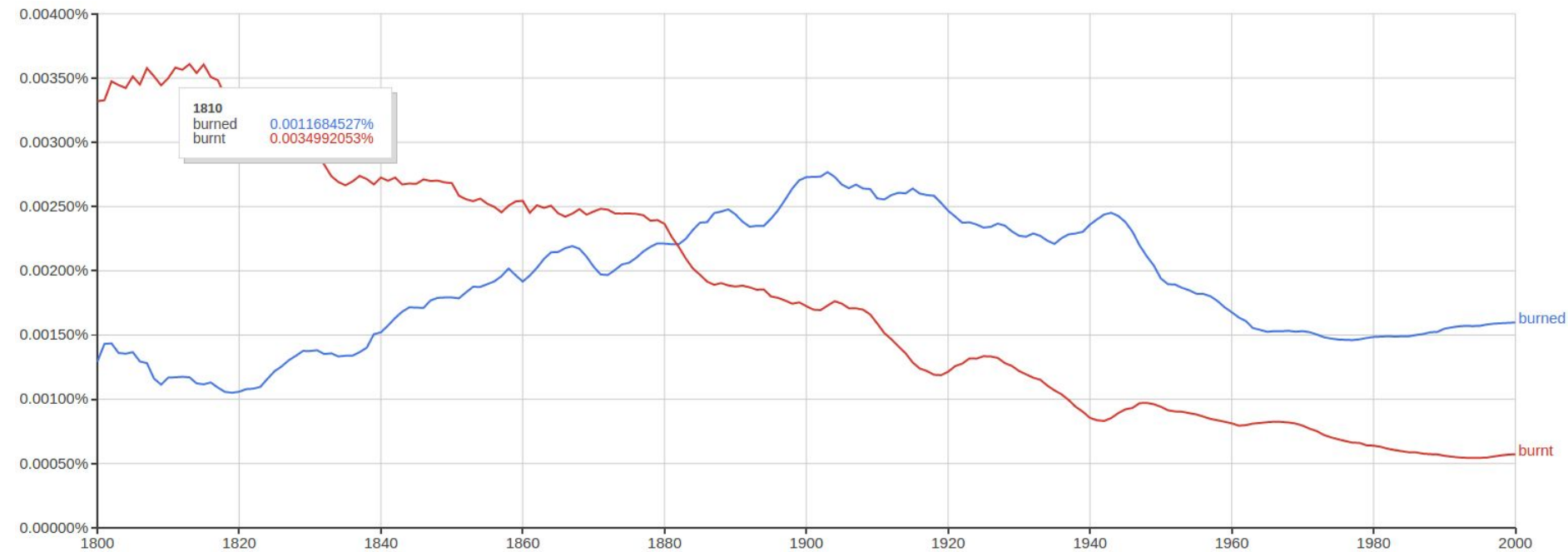


# CULTUROMICS

**A form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts.**

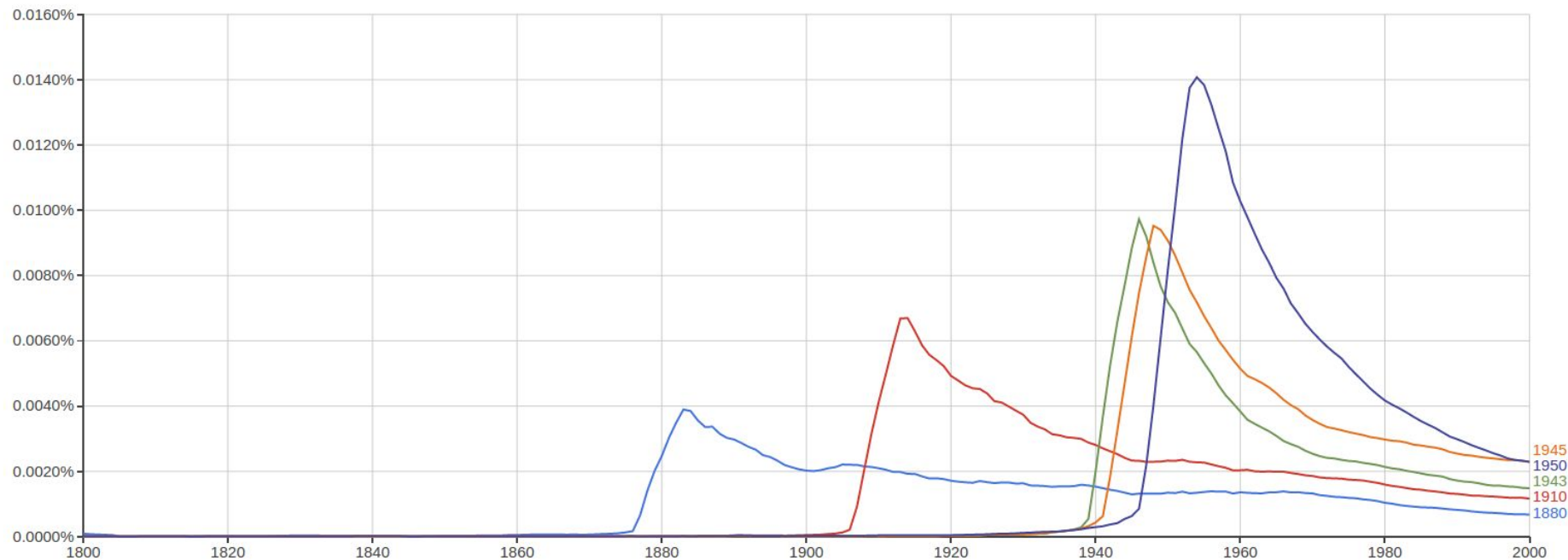
J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak e E. L. Aiden, *Quantitative Analysis of Culture Using Millions of Digitized Books*

# Culturomics: Grammar Evolution

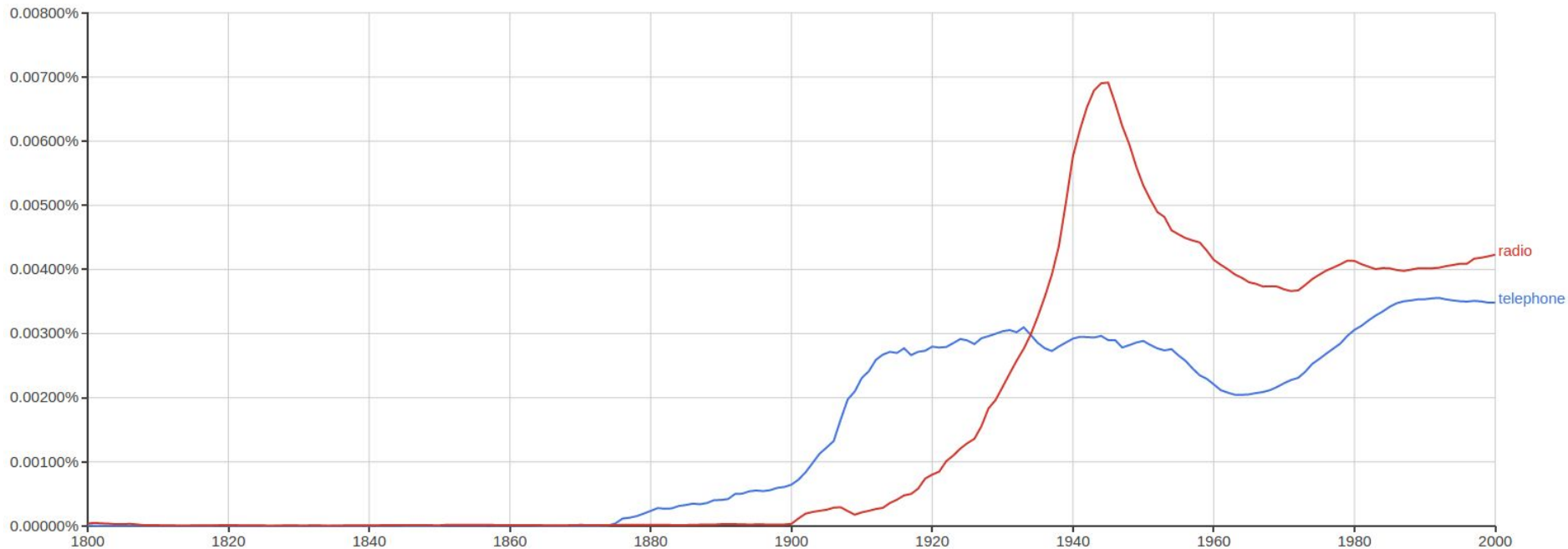




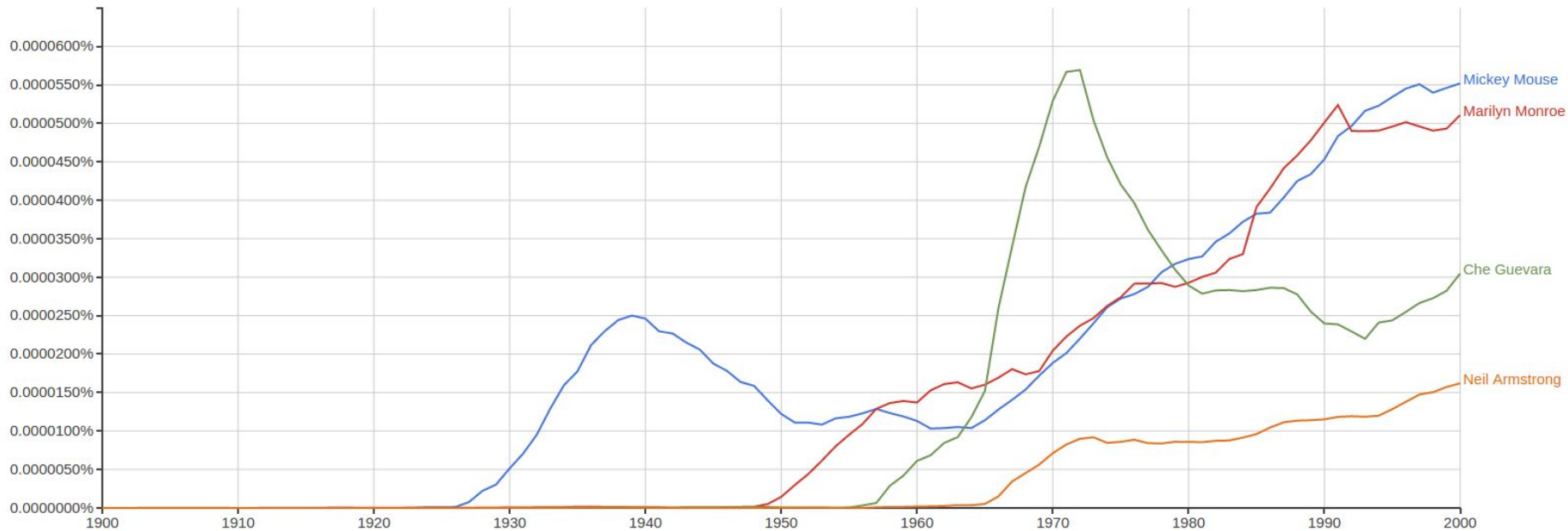
# Culturomics: Forgot the Old



# Culturomics: Forgot the Old

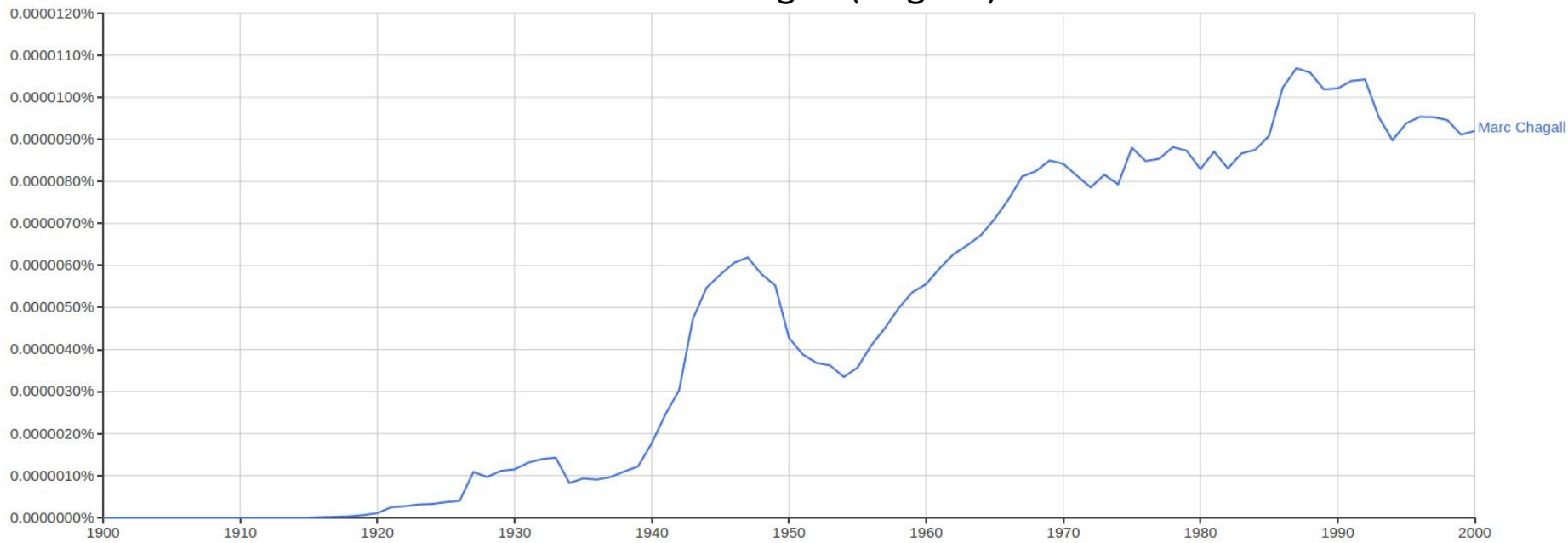


# Culturomics: Popularity



# Culturomics: Censorship

## Marc Chagall (English)

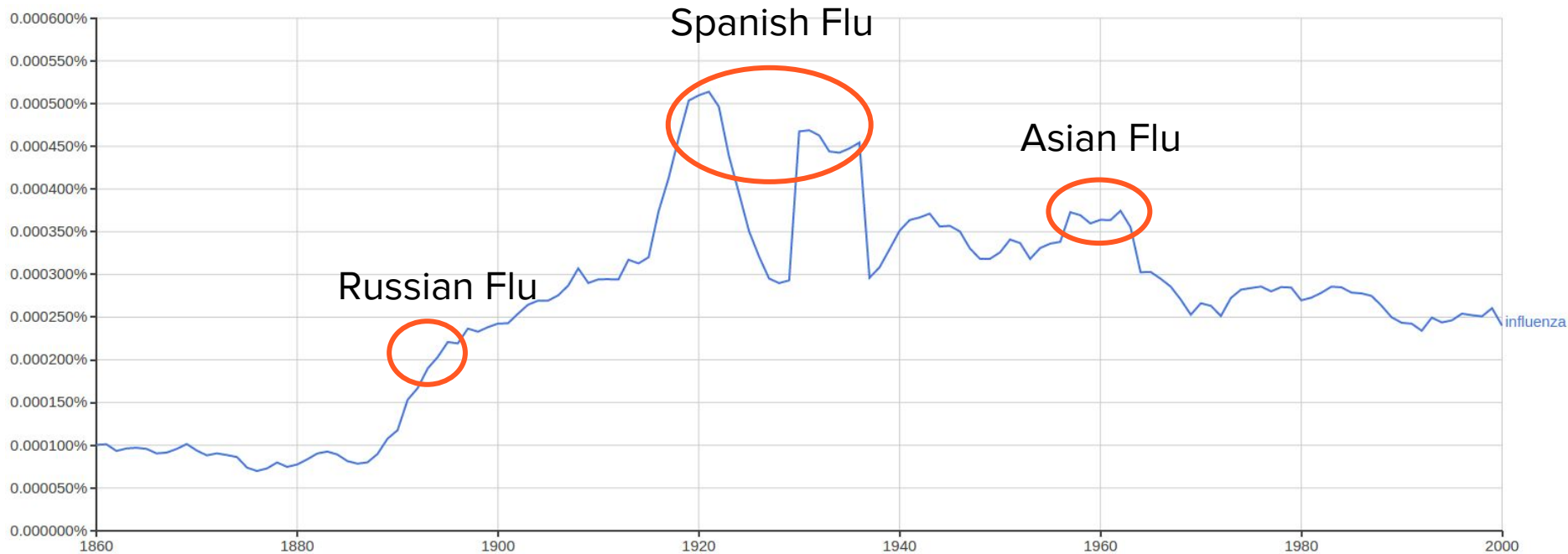


# Culturomics: Censorship

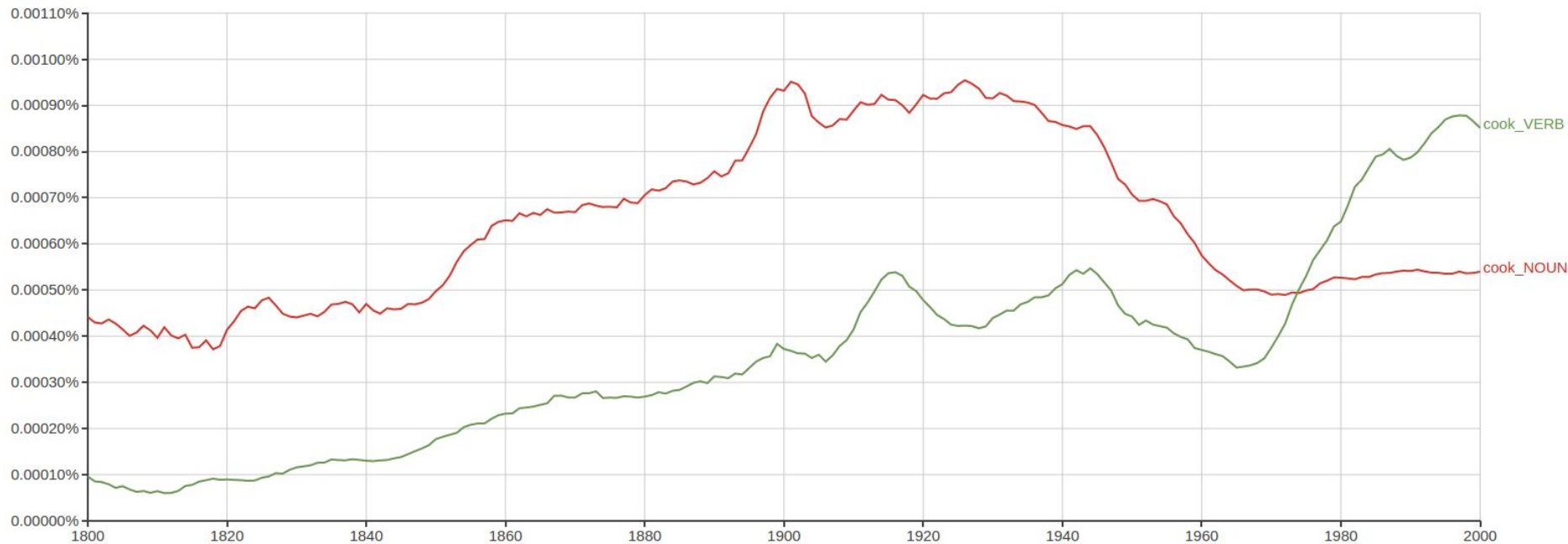
Marc Chagall (German)



# Culturomics: Events



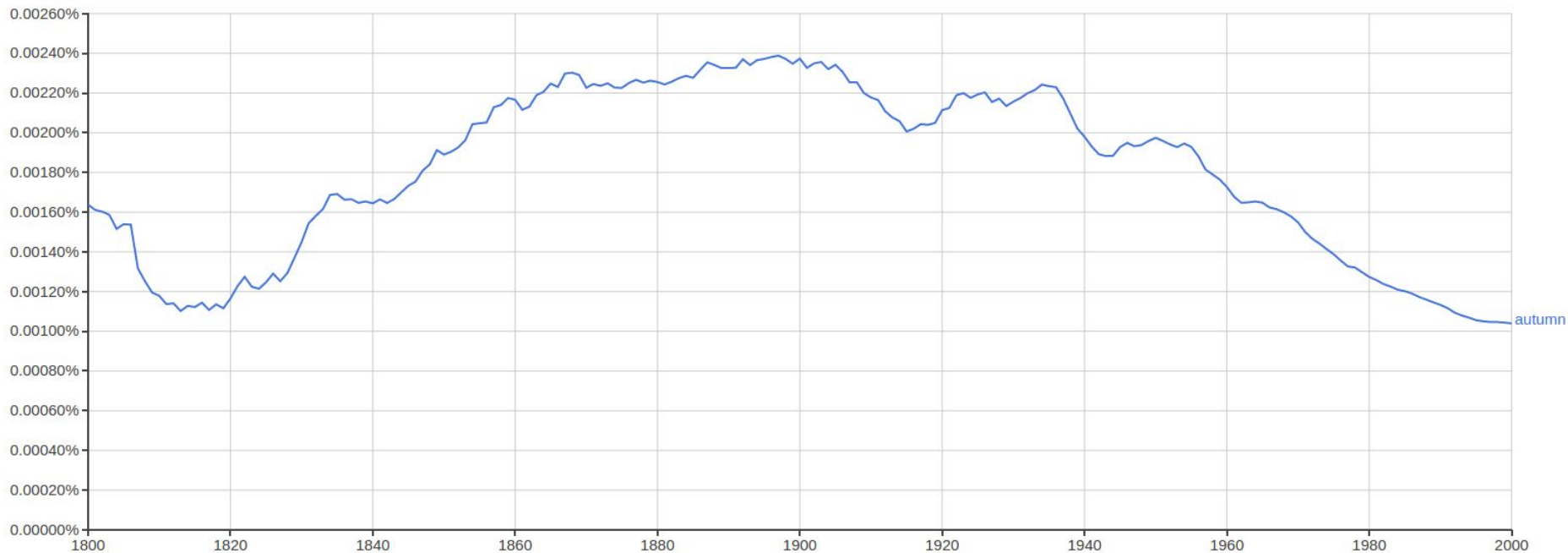
# Google N-gram Viewer: Part-of-Speech



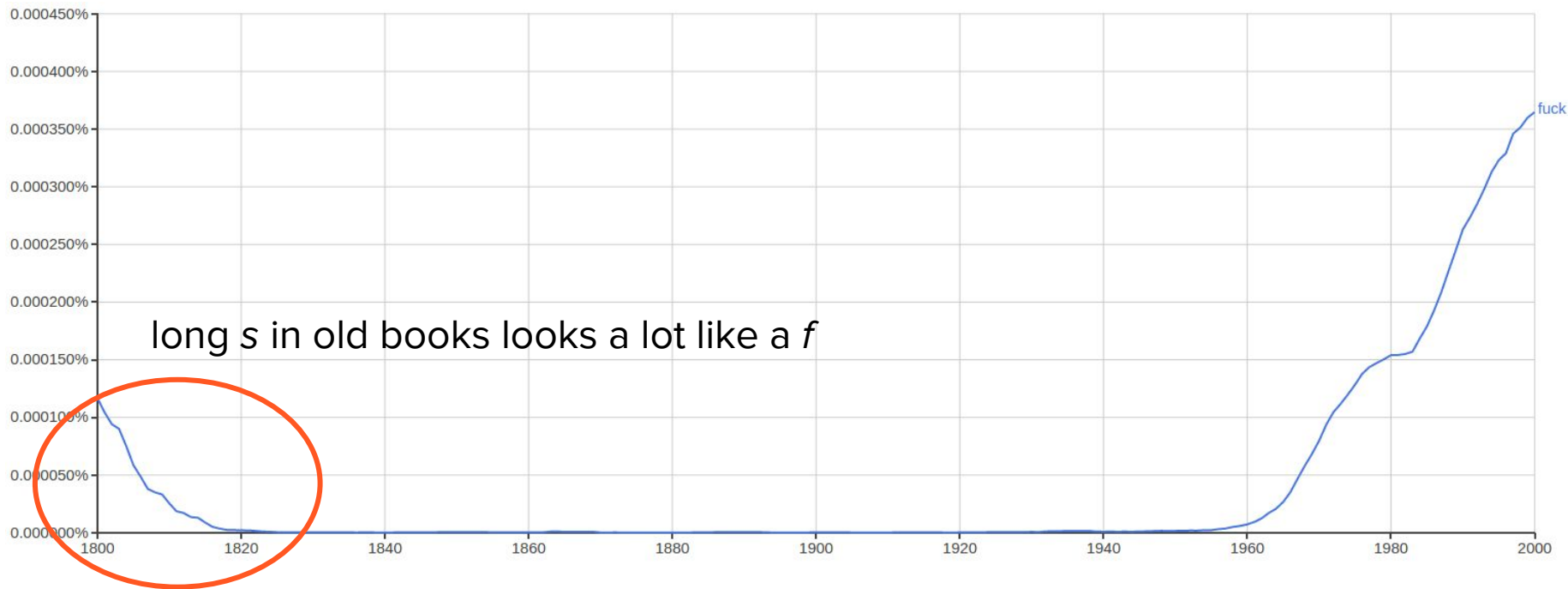
# Google N-gram Issues



# Overabundance of Scientific Literature

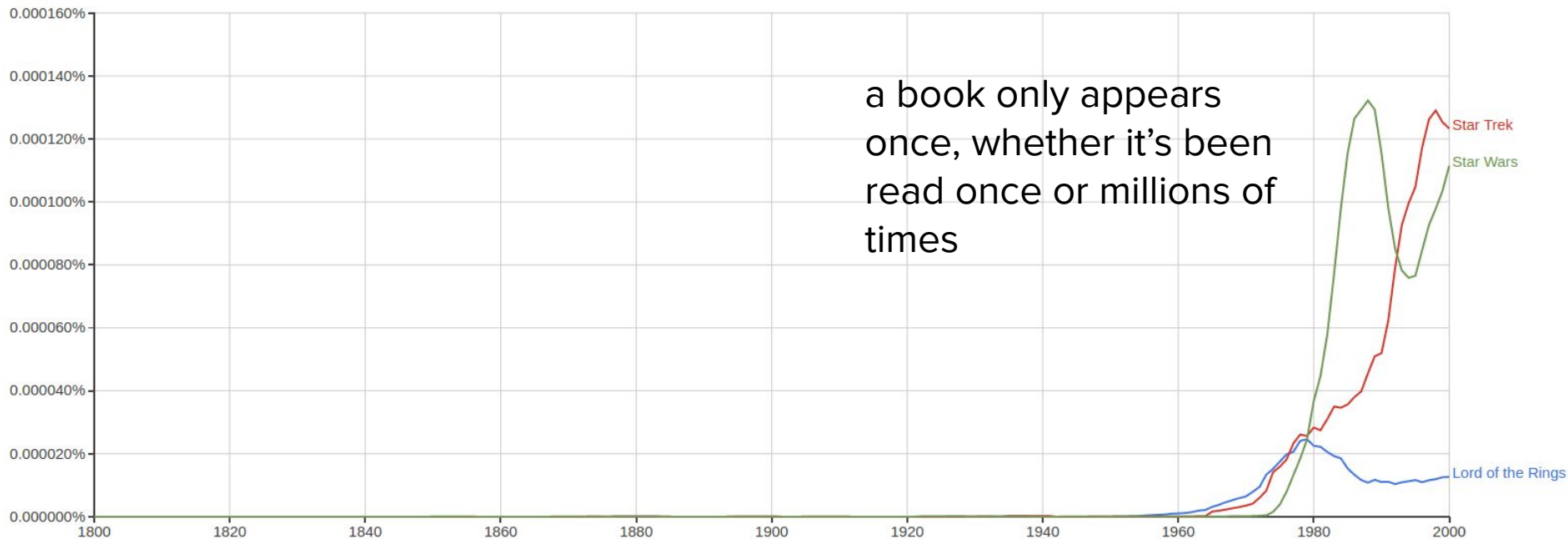


# OCR Errors



# Popularity Contests

a book only appears  
once, whether it's been  
read once or millions of  
times



**DUKweb,  
diachronic word  
representations  
from the UK Web  
Archive corpus**

# UK Internet Web Archive

- UK Web Archive collects, makes accessible and preserves web resources of scholarly and cultural importance from the UK domain
- **JISC UK Web Domain Dataset (1996-2013)**
  - resources from the **Internet Archive** that were hosted on domains ending in ‘.uk’

# DUKweb: source data

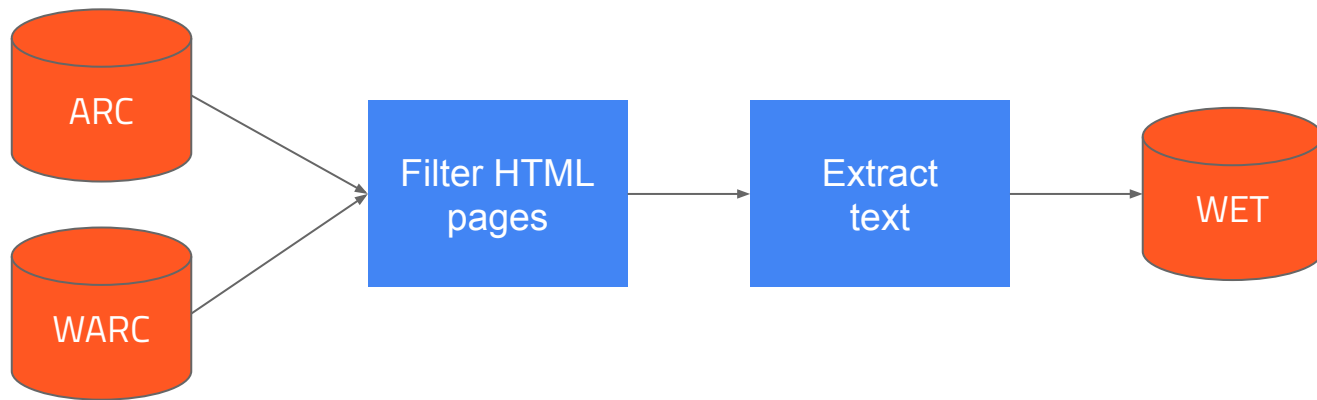
- **ARC format:** used to store "web crawls" as sequences of content blocks
- **WARC format:** is an enhancement of ARC format for supporting metadata, duplicate detection events and more

```
Content-Length: 7593  
  
HTTP/1.1 200 OK  
Server: Apache  
Content-Type: text/plain  
Date: Tue, 28 Dec 2010 02:32:55  
Keep-Alive: timeout=4, max=186  
Accept-Ranges: bytes  
Connection: close  
Last-Modified: Fri, 24 Dec 2010 11:11:11  
Content-Length: 7593  
  
User-agent: Googlebot  
Disallow: /iplayer/episode/*?from=r*  
Disallow: /iplayer/cy/episode/*?from=r*  
Disallow: /iplayer/gd/episode/*?from=r*  
Sitemap: http://www.bbc.co.uk/news_sitemap.xml  
Sitemap: http://www.bbc.co.uk/video_sitemap.xml  
Sitemap: http://www.bbc.co.uk/sitemap.xml  
Disallow: /cgi-bin
```

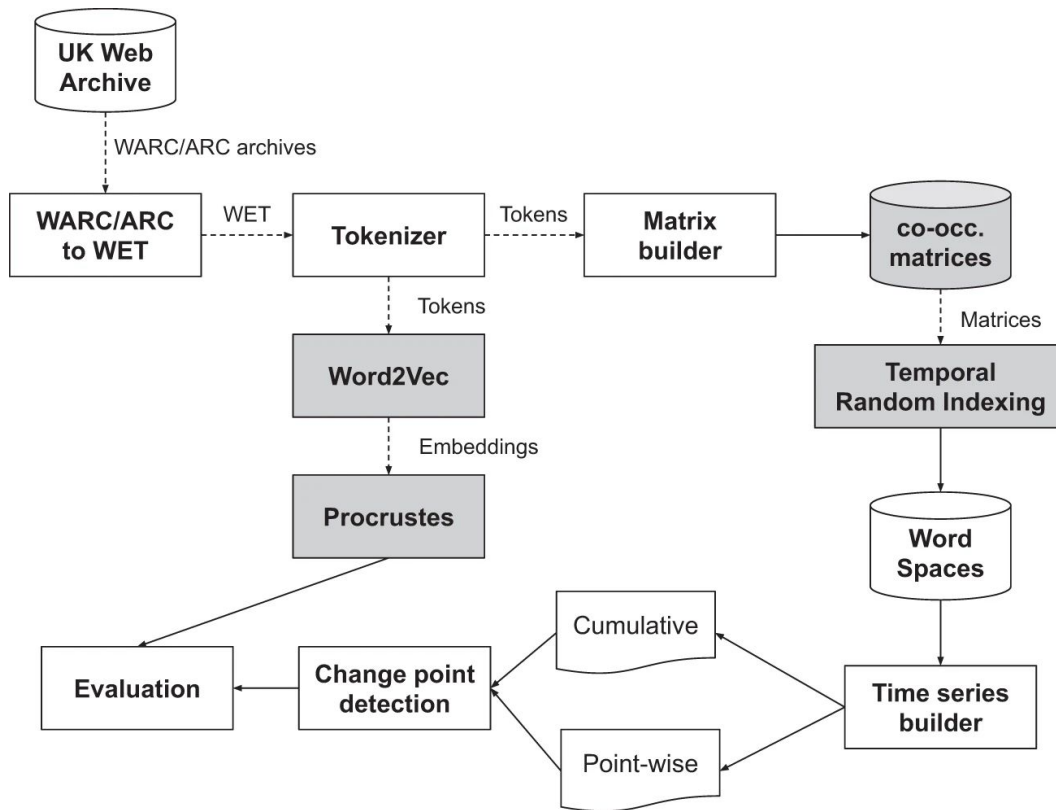
It is necessary to extract the textual content from HTML pages and discard all other types of content

# From ARC/WARC to WET

**WET format:** contains extracted plaintext from the data stored in ARC/WARC archives



# The creation of DUKweb





# Co-occurrence matrices

D-2012\_merge\_occ.gz

linux	swapping	4	google	173	xp	454	manufacturer	
237	job	64	install	255	security	137	cgi	47
operating	705	host	69	performance	44	sharing		
56...								

One matrix for each year

# Co-occurrence matrices

Target word

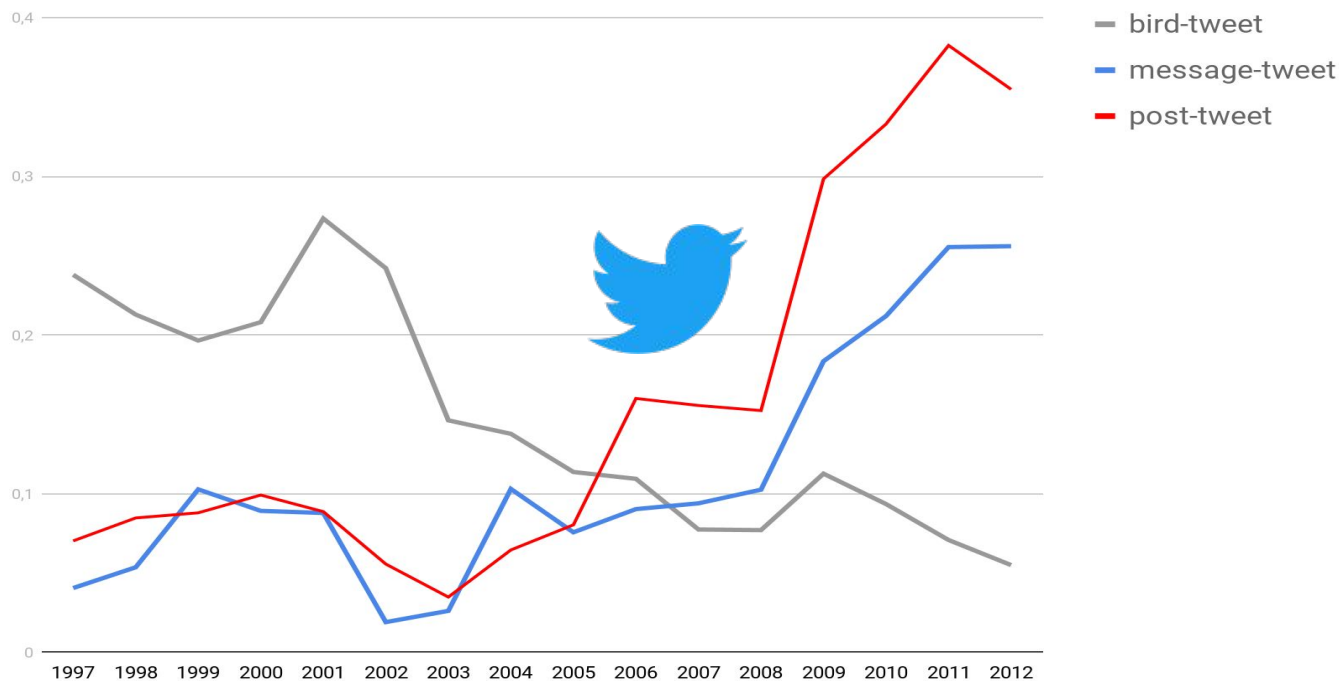
<b>linux</b>	swapping	4	google	173	xp	454	manufacturer	
237	job	64	install	255	security	137	cgi	47
operating	705	host	69	performance	44	sharing		
56...								

# Co-occurrence matrices

	co-occ. word	count						
linux	google	173	xp	454	manufacturer	237	job	
64	install	255	security	137	cgi	47	operating	
705	host	69	performance	44	sharing	56...		

co-occurrence

# Word Embeddings



# Available data

- **Co-occurrence matrices** for each year
- **Word2vec** aligned embeddings
- **Temporal Random Indexing** embeddings
- Download: <https://doi.org/10.23636/1209>

Adam Tsakalidis, Pierpaolo Basile, Marya Bazzi, Mihai Cucuringu & Barbara McGillivray. DUKweb, diachronic word representations from the UK Web Archive corpus. Nature Scientific Data  
<https://www.nature.com/articles/s41597-021-01047-x>

**Data**

**Annotation of Lexical  
Semantic Changes**

# Synchronic Word Sense Annotation

Strategies for the annotation of diachronic semantic change are similar to those applied for the annotation of word senses and polysemy at the synchronic level (see e.g. Erk et al 2013)

- 1) Annotation as classification of word occurrences with sense ID or dictionary senses.
- 2) Annotation as rating the applicability of dictionary senses on a graded scale to a word occurrence.
- 3) Annotation as rating the similarity between pairs of usages of the same word, also on a graded scale.

# Annotation of Lexical Semantic Change

- Related to the task and to the evaluation measure (as we will see later)
- Relevant for both LSC detection and LSC discovery



# The DUREl framework

- A framework that extends synchronic polysemy annotation to diachronic changes in lexical meaning.
- Use of a scale of semantic proximity among sentences (as previously done for synchronic research on polysemy, e.g. Soares da Silva, 1992; Brown, 2008; Erk et al., 2013)

Schlechtweg, D., Walde, S. S. I., & Eckmann, S. (2018).

Diachronic usage relatedness (DUREl): A framework for the annotation of lexical semantic change.

*arXiv preprint arXiv:1804.06517.*

# The DUREl framework

↑  
Identity  
Context Variance  
Polysemy  
Homonymy

↑  
4: Identical  
3: Closely Related  
2: Distantly Related  
1: Unrelated

Table 4: Blank (1997)'s continuum of semantic proximity (left) and the DUREl relatedness scale derived from it (right).

# The DUREl framework

- Given sentences of a target word  $w$  from two time periods  $t_1$  and  $t_2$ , the semantic relatedness of pair of sentences in each time period is annotated using the relatedness scale above.
- Low mean proximity in a period indicates polysemy or homonymy, while high mean proximity indicates meaning identity.
- The mean proximity values of the two time periods are compared: decrease or increase in the mean relatedness value of  $w$  from  $t_1$  to  $t_2$  indicate respectively innovative or reductive meaning change.

# The DUREl framework

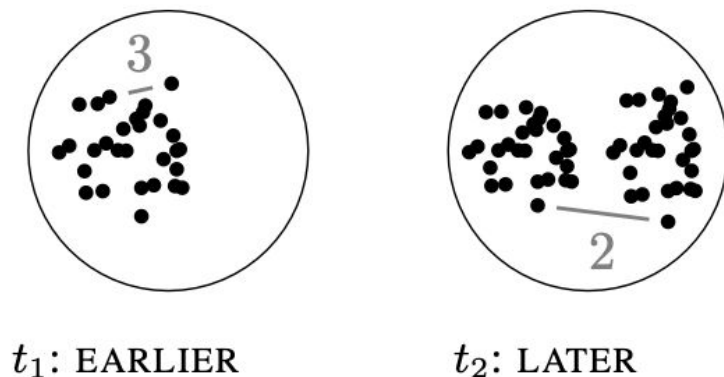


Figure 1: Two-dimensional use spaces (Tuggy, 1993; Zlatev, 2003) in two time periods with a target word  $w$  undergoing innovative meaning change. Dots represent uses of  $w$ . Spatial proximity of two uses means high relatedness.

# SemEval 2020 Latin annotation framework

- Annotators judged the relatedness between a use of a word  $w$  and a sense definition from a dictionary on the relatedness scale above

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

# DIACR-Ita 2020 framework

- Annotators were asked to assign each occurrence to one of the meaning of the lemma according to those reported in the Sabatini-Coletti dictionary

Basile, P., Caputo, A., Caselli, T., Cassotti, P., Varvara, R. (2020).  
DIACR-ITA @ EVALITA2020: Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task.  
*In Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).*  
Accademia University Press, Torino. ISBN: 9791280136275